

High-Dimensional Density Estimation via SCA: An Example in the Modelling of Hurricane Tracks[☆]

Susan M. Buchman, Ann B. Lee¹, Chad M. Schafer

Department of Statistics, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213

Abstract

We present nonparametric techniques for constructing and verifying density estimates from high-dimensional data whose irregular dependence structure cannot be modelled by parametric multivariate distributions. A low-dimensional representation of the data is critical in such situations because of the curse of dimensionality. Our proposed methodology consists of three main parts: (1) data reparameterization via dimensionality reduction, wherein the data are mapped into a space where standard techniques can be used for density estimation and simulation; (2) inverse mapping, in which simulated points are mapped back to the high-dimensional input space; and (3) verification, in which the quality of the estimate is assessed by comparing simulated samples with the observed data. These approaches are illustrated via an exploration of the spatial variability of tropical cyclones in the North Atlantic; each datum in this case is an entire hurricane trajectory. We conclude the paper with a discussion of extending the methods to model the relationship between TC variability and climatic variables.

Key words: dimension reduction, nonparametric density estimation, application to physical sciences

1. Introduction

In the realm of high-dimensional statistics, regression and classification have received much attention, while density estimation has lagged behind. Yet, there are compelling scientific questions which can only be addressed via density estimation using high-dimensional data and sampling from such estimates. Consider the paths of North Atlantic tropical cyclones (TC), some of which are shown in Figure 1. Temporarily assuming that tropical cyclones are independent and identically distributed, how would one use this data to estimate the probability that the next TC will make landfall at a particular swath of coastal North Carolina? Or how can one relate changes in TC paths over time to major climatic predictors such as sea surface temperature? If we cast each track as a single high-dimensional data point, density estimation allows us to answer such questions via integration or Monte Carlo methods.

All attempts to perform high-dimensional density estimation (HDDE) will require an element of dimensionality reduction to be feasible. Most existing methods, however, suffer from assumptions that are not appropriate for the applications presented above. Linear methods, such as Principal Component Analysis (PCA)[1], simply project all data points onto a lower-dimensional hyperplane, and are hence not able to describe complex, nonlinear variations. More recent work in HDDE has assumed sparsity of the input data [2], in the sense that the complex variations in density are a function of only a few of the original dimensions used to represent a datum. This is not typical of the data we consider here. For example, suppose one assumed that the density of a TC could be described by, say, three points on its path. This is conceding the ability to answer the questions described above, as we would have no information about the behavior of the track between these three points on the path.

Thus, there is a need for research on methods for nonparametric, nonlinear HDDE that involves dimensionality reduction and yet allows sampling from the original input space. We present an approach which utilizes a *spectral connectivity analysis (SCA)* method [3] called diffusion maps. SCA reparameterizes the data in a way that preserves context-dependent similarity. SCA and other eigenmap methods

¹Corresponding author. Email: annlee@stat.cmu.edu; tel.: +1 412 268 7831; fax: +1 412 268 7828

[☆]This work supported by ONR Grant N00014-08-1-0673

have been very successful for data parameterization [4, 5, 6], regression [7], and clustering and classification [8, 9, 5, 10]. In this work we extend these successes to a high-dimensional density estimator with the potential to address key questions regarding TC behavior, among other applications.

In what follows, we introduce a new approach to high-dimensional density estimation and sampling. The method is illustrated via the analysis of TC data. The basic idea of our approach is to perform density estimation in a reduced space using standard methods, to then generate a random sample from the lower-dimensional density, and finally to map the sampled data points back to the input space. We also describe a statistical method for evaluating the quality of the output of our simulation algorithm. While traditional techniques such as Q-Q plots and Kolmogorov-Smirnov tests work in very low-dimensional spaces, one needs to confirm that not too much information was lost in the reduction and that the original data set and the sample generated by the algorithm can reasonably be thought to come from the same distribution. We present a simulation-based approach that is independent of the particular HDDE method.

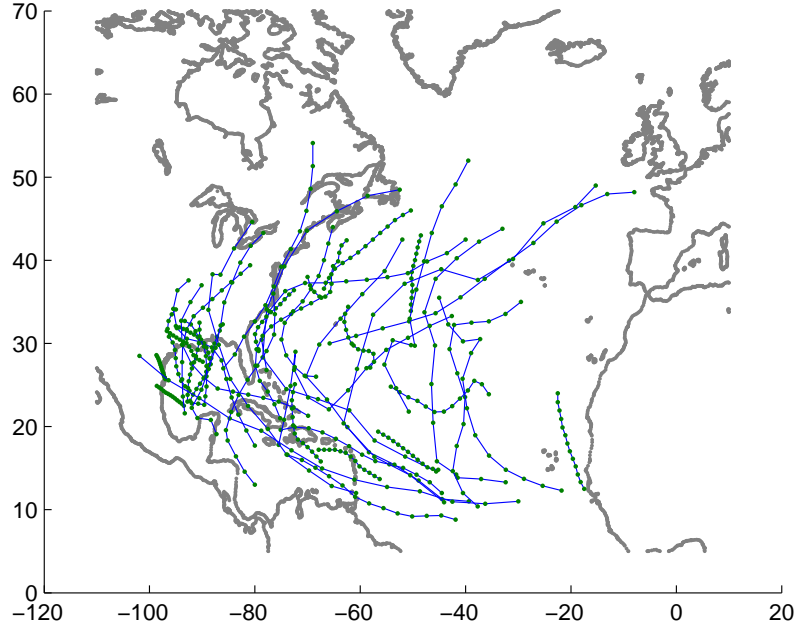


Figure 1: Forty randomly selected tracks of the 608 TCs between 1950 and 2006.

2. Methodology: Motivation and Description

The low-frequency, high-severity nature of tropical cyclones in the North Atlantic Ocean means that important and costly public policy, military, and business decisions are being made on the basis of relatively little historical data, and consequently any methodology that can extract more information from the data is valuable in advancing scientific, security, and economic interests. Much attention has been paid to hypotheses about the effect of various climatic predictors on TC occurrence frequency, TC landfall frequency, and TC intensity. However, few people have exploited the relationship between climatic predictors and the spatial variation of TCs, i.e. the TC tracks. This is largely due to the challenging nature of characterizing these relationships, and not due to a lack of importance. As Xie et al. [11] state, in addition to the focus on yearly counts and intensity, “it would be of great benefit to society if the preferred paths of hurricanes could also be predicted in advance of the onset of hurricane season.” Figure 1 illustrates the paths of 40 randomly selected tracks out of the 608 TCs that occurred between 1950 and 2006 [12].

To the extent that the spatial distribution of TC tracks has been investigated, researchers have primarily focused on variability in landfall location. For example, Hall and Jewson [13] address the

question of the effect of SST on landfall rates over fairly large regions of coastline; they use a rough-grained conditioning scheme which buckets years into “hot years” and “cold years”. Xie et al. [11] extend beyond landfall considerations and use empirical orthogonal functions to correlate climatic predictors with a “hurricane track density function” (HTDF). However, HTDF is somewhat of a misnomer, as the object they construct is not a density over tracks but a density over the ocean: the magnitude of the HTDF at x corresponds to x ’s proximity to observed hurricane tracks. The HTDF is really a reduction of the density; if one has a density over all tracks, one can construct the probability of a TC passing by a particular point by integrating the probability over all tracks which pass by that point, or via simulation.

The majority of work in track density estimation has adopted a similar approach, working in two spatial dimensions: first estimate a genesis (origination) density over the region of interest (e.g. the North Atlantic); then estimate a series of Markovian densities of track propagation, usually corresponding to 6-hour steps in which the distribution of the next location is a function of only the previous location; finally couple this with a lysis (death) component so that the simulated hurricane eventually stops [14, 15, 16, 17]. For example, Vickery et al. [17] uses the following model for changes in translation speed c and direction θ of a TC from time i to $i + 1$:

$$\begin{aligned}\Delta \ln c &= a_1 + a_2\psi + a_3\lambda + a_4 \ln c_i + a_5\theta_i + \epsilon \\ \Delta \theta &= b_1 + b_2\psi + b_3\lambda + b_4c_i + b_5\theta_i + b_6\theta_{i-1} + \delta\end{aligned}$$

where ψ and λ are latitude and longitude and ϵ and δ are error terms. In addition, to model spatial variability the parameters $a_1, a_2, \dots, a_5, b_1, b_2, \dots, b_6$ vary over *each box* in a $5^\circ \times 5^\circ$ grid over the Atlantic Ocean. Clearly, a primary drawback to this approach is the proliferation of parameters to estimate and models to validate.

To ground the explanation of the general methodology laid out in this paper, we will present its application to the task of estimating the density of the 608 TC tracks between 1950 and 2006, forty of which are shown in Figure 1.

2.1. Dimensionality Reduction

The first step in our algorithm is to perform dimensionality reduction and reparameterize the data using a SCA technique called diffusion maps [4, 5]. Diffusion maps rely on a metric that quantifies the “connectivity” of a data set and introduces a new coordinate system based on this metric.

Diffusion Maps

The construction begins by creating a weighted graph $G = (\Omega, W)$ where the nodes in the graph are the n observed data points in \mathbb{R}^d , e.g. the trajectories of TCs discretized to d spatial locations. The weight given to the edge connecting two data points $x \in \Omega$ and $y \in \Omega$ is $w(x, y) = \exp(-\Delta^2(x, y)/\epsilon)$, where $\Delta(x, y)$ is an application-specific locally relevant distance measure and ϵ controls the neighborhood size. We construct a Markov random walk on the graph where the probability of stepping directly from x to y is $p_1(x, y) = w(x, y) / \sum_z w(x, z)$. This probability will be small unless the two points (such as two TCs) are similar to each other. We then iterate the walk for t steps, and consider $p_t(x, \cdot)$, the conditional distribution after t steps having started at x . One natural way to think of two points as similar is if their t -step conditional distributions are close; formally, we define the “diffusion distance at scale t ” as:

$$D_t^2(x, y) = \sum_z \frac{(p_t(x, z) - p_t(y, z))^2}{\phi_0(z)}. \quad (1)$$

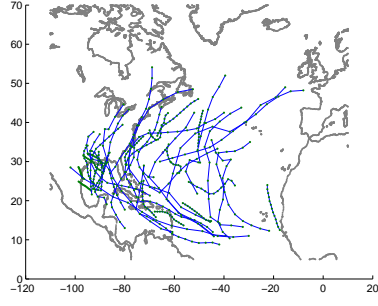
The stationary distribution $\phi_0(\cdot)$ of the random walk penalizes discrepancies on domains of low density.

One can show there is a mapping — the diffusion map — which reduces the dimension while still approximating diffusion distance. Each high-dimensional data point x can be mapped to reduced dimension m via the following transformation:

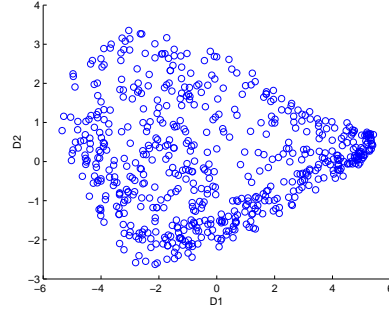
$$\Psi_t : x \mapsto [\lambda_1^t \psi_1(x), \lambda_2^t \psi_2(x), \dots, \lambda_m^t \psi_m(x)] \quad (2)$$

where λ_j and ψ_j represent the eigenvalues and right eigenvectors of $P = \{p(x, y)\}_{x, y}$, the row-normalized similarity matrix. One can show that

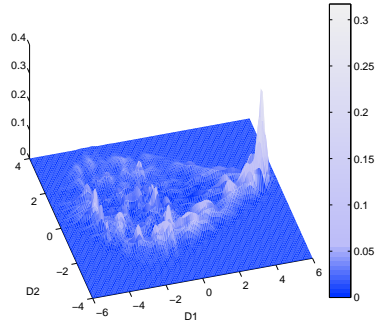
$$D_t^2(x, y) = \sum_{j=1}^{n-1} \lambda_j^{2t} (\psi_j(x) - \psi_j(y))^2 \approx \sum_{j=1}^m \lambda_j^{2t} (\psi_j(x) - \psi_j(y))^2 = \|\Psi_t(x) - \Psi_t(y)\|^2. \quad (3)$$



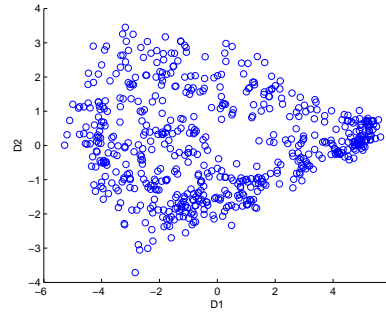
(a) A subset of the observed tracks.



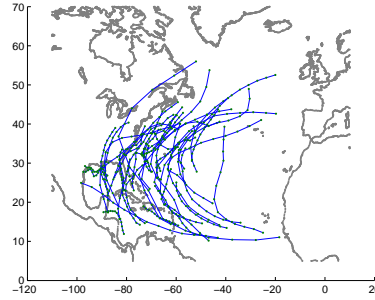
(b) The observed tracks in diffusion space.



(c) A density over diffusion space.



(d) A random sample from the density.



(e) A subset of the sample mapped back into track space.

Figure 2: **An overview of the dimensionality-reduction approach to TC track simulation.** (a) shows 40 randomly selected tracks out of a total of 608 TCs observed between 1950 and 2006. (b) shows all 608 tracks mapped to diffusion space for $m = 2$ and $\epsilon = 430$, with each point corresponding to a particular track in (a). (Although the analysis in this paper works with $m = 3$, we use the two-dimensional map here to be able to visualize the whole process.) An estimated density for the diffusion space data of (b) is visualized in (c), and a 608-element sample from that density is shown in (d). Each point in the sample can be interpreted as being associated with a new, as-yet-unobserved track. The sample is finally mapped back into track space; 40 randomly selected TCs of the sample are shown in (e).

In other words, the Euclidean distance between two nodes in the reduced space approximates their diffusion distance, which in turn reflects the *connectivity* of the data as defined by a t -step random walk. Richards et al. [7] demonstrate that the diffusion coordinates can be used to meaningfully model physical information in a regression setting.

Here we focus on pure spatial similarity: Do the TCs follow similar paths? We begin by regularizing each track, transforming it from its original representation with a varying number of points, which represent the position of the TC in six-hour increments, to a new representation with thirteen regularly-spaced points. We define the distance measure $\Delta(x, y)$ to be the sum of the Euclidean distance between the 13 corresponding points on each track pair (see Figure 2.1) and we use the dimension $m = 3$, the selection of which is described at the end of Section 2.3. We stress that $\Delta(\cdot, \cdot)$ need not be Euclidean distance; we can use an application-specific distance measure. One could, for example, imagine a different project in which two TCs are considered to be similar not only if their tracks are spatially similar but also if their intensity histories are similar.

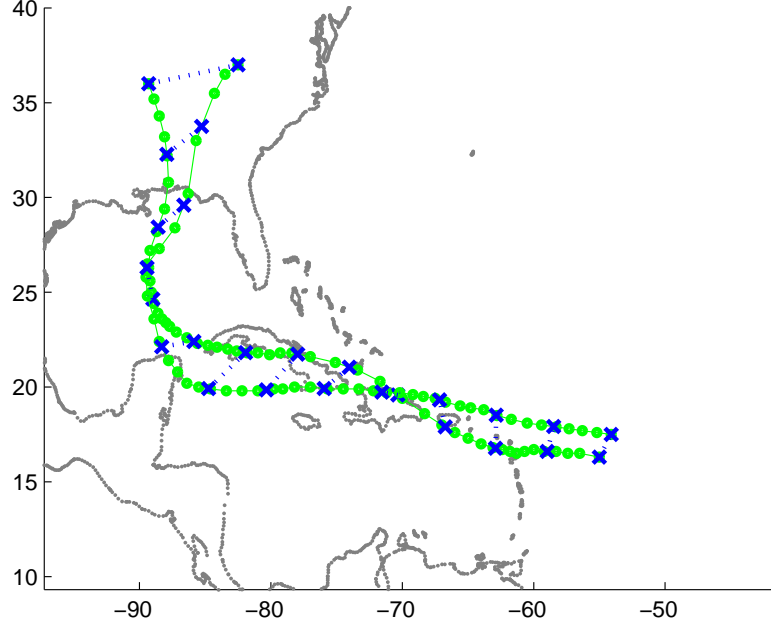


Figure 3: Two regularized tracks are shown above: the original 6-hour segments are marked by circles, and the regularized segments are marked by \times . The application-specific distance measure is the sum of the distances between 13 corresponding pairs of points, shown as dashed lines.

The issue of how to select the parameters $\Theta = (\epsilon, t)$ in SCA is usually an open problem. Below we present an approach that is closely connected to the so-called Nyström extension of the map.

Nyström Extension and Parameter Selection

As the diffusion map is defined only on the points in the graph, we need a technique to project new points into the map. In other words, we wish to extend ψ_i , the right eigenvectors of the transition matrix, to a new value y . We know that for $\forall x_1, x_2 \in \Omega$,

$$\sum_{x_2 \in \Omega} p(x_1, x_2) \psi_j(x_2) = \lambda_j \psi_j(x_1). \quad (4)$$

Hence, we can approximate an extension to y by replacing x_1 with y :

$$\frac{1}{\lambda_j} \sum_{x_2 \in \Omega} p(y, x_2) \psi_j(x_2) = \hat{\psi}_j(y). \quad (5)$$

This leads to the extended diffusion map

$$\widehat{\Psi}_t : y \mapsto \begin{pmatrix} \lambda_1^{(t-1)} \sum_{x_2 \in \Omega} p(y, x_2) \psi_1(x_2) \\ \lambda_2^{(t-1)} \sum_{x_2 \in \Omega} p(y, x_2) \psi_2(x_2) \\ \vdots \\ \lambda_{q(t)}^{(t-1)} \sum_{x_2 \in \Omega} p(y, x_2) \psi_{q(t)}(x_2) \end{pmatrix}. \quad (6)$$

We also need to extend the transition matrix to have a y -row, but this can be done exactly: simply apply the distance function to y and all members of Ω , and then normalize the row.

With the extension, we can now perform cross-validation of a collection of parameters, Θ :

1. Hold out observation x_i from the test set.
2. Compute $\Psi_{(-i); \Theta}$, the diffusion map computed with all observations but x_i .
3. Use the Nyström extension to find $\widehat{\Psi}_{(-i); \Theta}(x_i)$, the projection of x_i into diffusion space.
4. Find $\widehat{x}_i = \Psi_{(-i); \Theta}^{-1}(\widehat{\Psi}_{(-i); \Theta}(x_i))$, the pre-image of the projection of x_i .
5. Calculate $\Delta(x_i, \widehat{x}_i)$, the distance between the true track and its approximation.
6. Repeat for all observations, returning $\sum_i \Delta(x_i, \widehat{x}_i)$ as the error for Θ .
7. Repeat for all candidate values of Θ .
8. Return as the model parameters the value of Θ whose error is smallest.

Given the Nyström extension, all the steps are straight-forward except for step 4. How do we find pre-images of points in diffusion space? As this is an important step later in the algorithm as well, let us consider it more generally: let ζ be the point in diffusion space we want to find a pre-image for. As emphasized in Arias et al. [18], we search for the point in the original space whose extension into diffusion space comes closest to ζ in Euclidean distance. In other words, we want to select

$$\widehat{\Psi}^{-1}(\zeta) = \arg \min_{x \in \mathbb{R}^d} \|\zeta - \widehat{\Psi}(x)\|^2. \quad (7)$$

The implementation of the pre-image search will be context-specific; we defer the discussion of the pre-image search for the TC density until Section 2.2. Figure 2(b) shows a two-dimensional diffusion map of the TC data.

Density Estimation and Random Sampling

Once the data are mapped into the low-dimensional space, a range of nonparametric density estimators could be employed. In our initial work, we use k-nearest-neighbor kernel density estimation (rather than density estimation with a fixed kernel bandwidth), because of the tendency for data points to cluster near an apparent “boundary” in diffusion space. Also, using this estimator, simulation is trivial [19]. Figures 2(c) and 2(d) show an example of density estimation and random sampling for TC data. Future work will focus on fitting models which allow for natural incorporation of covariates and time dependence; see Equation 11 in Section 3.

2.2. Inverse Mapping

In order to verify the validity of our density estimate, and also to be able to simulate new sample tracks, we need a method for finding the pre-image of an arbitrary point in diffusion space. As mentioned, using the Nyström extension and pre-image objective of Equation 7 there is a natural approach: the pre-image can be approximated as the track whose projection into diffusion space comes closest to the point that we wish to invert. In practice, however, designing a search mechanism that is both sufficiently exhaustive and computationally feasible is difficult. One solution is to restrict the pre-image to be a convex combination of observed data objects, assuming they are of the same dimension (or can be approximated as such in a meaningful way) [20, 21]. This is the approach we describe here.

Let ζ be the point in diffusion space for which we seek the pre-image. The Euclidean distance from ζ to $\Psi(x)$ for each observed track $x \in \Omega$ is a natural measure of the similarity between $\Psi^{-1}(\zeta)$ and x . Thus, we can construct weights as

$$w(\zeta, x) = \frac{\exp(-\|\zeta - \Psi(x)\|^2 / \sigma_\zeta)}{\sum_{y \in \Omega} \exp(-\|\zeta - \Psi(y)\|^2 / \sigma_\zeta)},$$

and then use these weights in constructing the convex combination:

$$\hat{\Psi}^{-1}(\zeta) = \sum_{x \in \Omega} w(\zeta, x) x. \quad (8)$$

The problem has thus been reduced to searching over a single parameter σ_ζ which controls the spread of the normal kernel used to determine the weights. This approach does require that each track be condensed to an equal number of points along its path, but that number need not be small in order for this to be feasible. Furthermore, note that in practice it will typically only be necessary to interpolate between tracks which are similar, i.e. if $w(\zeta, x)$ and $w(\zeta, y)$ are both large, then x and y will usually be alike.

However, constructing inverses as convex combinations of observed tracks produces simulated tracks that are never more extreme than the most extreme observed track for whichever spatial measurement one might want to consider. For example, using this approach, no pre-image could be longer than the longest of the observed tracks or shorter than the shortest of observed tracks. To overcome this shortcoming, in addition to searching over σ_ζ , we also allow the average to stretch up to 150% and shrink down to 75%. In other words, we do not search for merely the convex combination whose projection comes closest to η , but we treat the convex combination as a form which can be stretched (in both directions) by the optimal factor in $[.75, 1.5]$. Furthermore, we consider separately a shrink/stretch anchored at the origination point and the lysis point of the convex combination. Despite these extensions, one remaining shortcoming is that it is not possible to sample a track with more loops than any of the observed tracks, although it is possible for such a TC to occur. A set of 608 tracks simulated using our algorithm is shown in Figure 2(e). Other variations on the pre-image construction can and will be explored.

More importantly, we are developing ways of testing the validity of the entire analysis pipeline, as summarized in Procedure 1. Our approach to assessment relies upon comparing the observed tracks to simulated tracks, which is described next.

Procedure 1 High-dimensional density estimation

Input: Ω , a high-dimensional data set of dimension d and size n ;

$\Delta : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$, a locally-relevant distance measure.

Output: $\hat{\Omega}$, a sample from the estimated density of Ω .

1: **Dimensionality reduction:**

2: Construct Ψ , an m -dimensional diffusion map for Ω, Δ :

3: Select model parameters ϵ and t via cross-validation.

4: Perform density estimation in diffusion space to form $\hat{\mu}$.

5: Generate Φ , a size- n random sample from $\hat{\mu}$.

6: **Inverse mapping:**

7: Find $\hat{\Omega} = \hat{\Psi}^{-1}(\Phi)$, the pre-image of Φ in the d -dimensional input space.

8: Validate results via repeated simulation from $\hat{\mu}$.

2.3. Validation

Although best efforts at verification were made in each stage of the above algorithm — cross-validating the diffusion map parameters, selecting reasonable bandwidths in the density estimation — we would like to have a procedure for making a comprehensive evaluation of the resulting estimate. Our approach will be based on a test of the hypothesis that two high-dimensional samples — the observed data and simulated data — come from the same underlying distribution.

We are pursuing results which will prove consistency of the density estimator, but this will be challenging, in part because it will require a new analysis every time a part of the algorithm is modified (for example, if one wanted to move from kernel density estimation to locally linear density estimation). Here we produce a nonparametric high-dimensional verification technique that treats the particulars of the methodology as a black box. We assess whether a new sample can reasonably be said to come from the same distribution as the observed data, regardless of how the former was generated. This is analogous to existing tools for one-dimensional analysis (Q-Q plots, the Wilcoxon rank-sum test, the two-sample Kolmogorov-Smirnov test). While there are multivariate extensions to some of these classic tests [22], these methods often struggle with extensions beyond two dimensions. We utilize a simple test statistic similar to that given in Hall and Tajvidi [23], which allows for genuine high-dimensional comparisons,

and also yields a visual assessment tool for helping to identify, and therefore possibly correct, the ways in which the simulation fails.

We make a connection between the choice of the local distance metric Δ and the validation of the density estimate; in practice, this connection can be used in motivating the choice of Δ . Formally, let μ_1 and μ_2 be two distributions over the input space and let X_1, X_2, \dots, X_n be i.i.d., distributed as μ_1 , and let Y_1, Y_2, \dots, Y_n be i.i.d., distributed as μ_2 . Define the quantity $\mathcal{L}_\Delta(\mu_1, \mu_2)$ to be the proportion of the values

$$(X_1, X_2, \dots, X_n, Y_1, Y_2, \dots, Y_n) \quad (9)$$

whose nearest neighbor (as measured by Δ) is from the same sample. Let $\mathcal{R}_\Delta(\mu_1, \mu_2) = \mathbb{E}(\mathcal{L}_\Delta(\mu_1, \mu_2))$. We define a density estimator to be *consistent with respect to local distance metric Δ* if

$$\lim_{n \rightarrow \infty} \mathcal{R}_\Delta(\hat{\mu}_n, \mu_X) = 0.5, \quad (10)$$

where $\hat{\mu}_n$ is the estimated distribution, and μ_X is the true distribution. Heuristically, if the two distributions $\hat{\mu}_n$ and μ_X are the same, then the nearest neighbor of any sample value is equally likely to be from either of the two samples.

A simulated test

In practice, how can one use the motivation behind the more formal notion of consistency with respect to the local distance metric to produce a test of our sampling mechanism? Noting that all samples generated by the algorithm will be from the same distribution, we can use a simulation-based approach:

1. For some large number k , generate k pairs of samples of size n using the algorithm.
2. For the i^{th} pair, calculate and record \mathcal{L}_Δ .
3. Generate one last sample of size n using the algorithm and pair it with the observed tracks; calculate $\ell^* = \mathcal{L}_\Delta$ for these values.
4. Evaluate where ℓ^* falls in the distribution of the k proportions; reject the hypothesis that the observed tracks come from the estimated density if it is too far in the tails.

This test can be adapted to any sampling mechanism, not just the one presented in this paper. Note that rejecting the null hypothesis is not always indicative of a problem. In addition to verifying a sampling method, this technique can be used to find and hone in on regions of dissimilarity among samples that one *expects* to differ – Section 3 provides an example.

When this test was applied to the observed tracks and the sample shown in Figure 2(e), there were 689 within-sample nearest neighbors, for a proportion of $\frac{689}{2 \cdot 608} = 0.56$. For $k = 1000$, there were 31 pairs whose within-sample NN proportion was higher, despite being from the same distribution. This indicates that there is room for improvement in the steps of our algorithm, but the simulated data are fairly similar to the observed data. Given that we used only $m = 3$, this is quite encouraging.

Visual assessment

In a Q-Q plot one can often immediately diagnose the nature of dissimilar samples (for example, one sample having heavier tails than another). However, as emphasized in Hall and Heckman [24], it is much harder to craft visualization methods for high-dimensional data, as they tend to be co-dimensional with the data. But if one of the samples is created via a method that involves dimensionality reduction, this can be used in conjunction with the local distance metric to provide a quick visual gauge of the region of dissimilarity.

We plot both the original data and the sampled data in diffusion space, distinguishing the points not based on which sample they came from but by whether their nearest neighbor is of the same or different sample. One might be able to visually identify a region in the diffusion map which is too saturated with data points who have within-sample nearest neighbors. Of course, this will not provide as easy an answer as “different thickness of tails” or other causes of one-dimensional dissimilarity, but by inspecting the data points in the saturated region in their higher-dimensional representation, it can provide one with tools for generating hypotheses on why the simulation is insufficient which a single test statistic would not be able to do.

Choice of dimension

Note that there is a tradeoff between the dimensionality reduction – in which a larger m improves the results by retaining more information – and the density estimation – in which a larger m makes density estimation more difficult. In selecting $m = 3$ we used a criterion that encompasses all components of the method (as opposed to selecting m at the same time that ϵ and t are estimated). In particular, to evaluate a fixed choice of m , proceed with Procedure 1; after performing the density estimation, generate 100 sets of simulated tracks of size $n = 608$. For O , the set of observed tracks, and S_i , the i^{th} set of simulated tracks, find $\mathcal{L}_\Delta(O, S_i)$. Average these 100 ratios. Select the dimension whose average ratio is closest to 0.5 as the optimal dimension. For this application we considered $m \in \{2, 3, 4\}$, this kernel density estimation in greater than four dimensions is not reliable with only 608 observations; in Section 3.1 we discuss a form of orthogonal series density estimation that we expect will perform well in higher dimensions.

3. Conditional Density Estimation

Some of the most important questions regarding TCs could be addressed, at least partially, through a better understanding of the relationship between TC occurrence and other measurable characteristics of the climate system. Such relationships could be utilized in, for instance, creating and verifying complex simulation models, predicting future trends in TC activity, and understanding human influence on the climate system. Specifically, an area of great concern is the effect that rising sea surface temperatures (SST) might have on the frequency and/or intensity of TCs. The simulation method introduced in this paper can be applied to the question of how changes in sea surface temperature might affect the landfall distribution of TCs.

First consider a set-up similar to Hall and Jewson [13]: they focused on the 19 hottest and 19 coldest years from 1950 to 2005, where a year’s “temperature” was defined as the July-August-September SST averaged over a region of the Atlantic. After dividing the North American continental coastline into six major segments — the U.S. Northeast, the U.S. mid-Atlantic, Florida, the U.S. Gulf, the Mexican Gulf, and the Yucatan peninsula — they performed hypothesis tests on the difference in yearly landfall rates between the hot and cold years. In all regions but the U.S. Northeast, the landfall rate was higher in hot years, with the difference in the Yucatan being found as statistically significant.

Their approach requires that one assume a particular theory about the relationship between SST and landfall rates. It also requires that the coastline be divided into somewhat arbitrary, large segments. It would be preferable to have the densities inform us as to the regions which are experiencing differences among the hot and cold years. For example, consider Figure 4, which shows the density estimate found when our methods are applied separately to the tracks from cold years (Figure 4(a)) and the tracks from hot years (Figure 4(b)). If we focus on the regions in which the hot density is much higher than the cold density — specifically, the range $D1 \in [2.4, 3]$, $D2 \in [-.9, -.4]$, found using the technique of Section 2.3 — we can map the tracks that fall into that region, as shown in Figures 4(c) and 4(d). We see that most of these tracks are southern U.S. and Central American landfalling tracks, which comports with Hall and Jewson [13]. Density estimation allows us not only to test hypotheses about the effect of climatic predictors on TCs, but also provides a way of generating insight into the nature of these relationships.

The above analysis illustrates how the methods described in this paper can be used to investigate questions about TC tracks, but it reduces SST to a binary variable: “hot year” or “cold year.” These results treat a 56-year stretch of extreme climate events as samples from two distributions. For those with long-term interests, such as insurers or bodies that establish coastal building codes, this may be sufficient, as they are interested in the distribution of extreme events over the next 50 years, but many important questions concern short-term temporal and spatial variation in TC distribution. Thus, our primary interest in the study of TC tracks is not in estimating a single high-dimensional density, but instead to quantify the changes in the track-space density over time, and to model how these changes relate to other climate variables.

In fact, SST data is available as a high-resolution time series, as a function of ocean position. We have begun to exploit these data when fitting models; the potential of such a model is illustrated in Figure 5. In this example, a three-dimensional diffusion map is created using 1000 TC tracks since 1900; this map is shown in the upper-left panel of the figure. Consider a point in three-dimensional diffusion space, $\mathbf{z}_0 = (0.39, 0.086, -0.0098)$. The upper-right panel of Figure 5 shows all of the tracks which are within a small diffusion distance (i.e. small Euclidean distance in diffusion space) of this point; these are

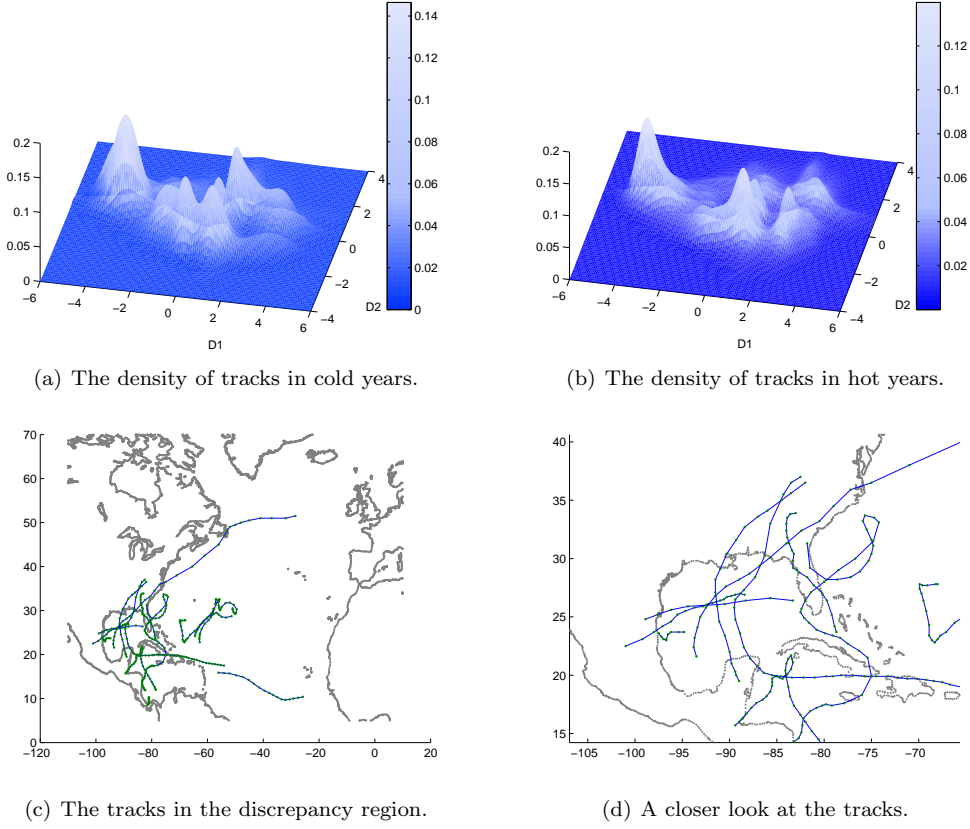


Figure 4: Densities for tracks conditioned on hot and cold years; the figures in the bottom row shows the tracks from the region where the density is much higher in the hot years.

a cluster of storms which remain far from the Atlantic coast. The dashed line in the lower-left panel shows the change in the density of tracks near \mathbf{z}_0 over all of the years, smoothed over time.

This panel also depicts the sea surface temperature at latitude/longitude $(30W, 15N)^2$, chosen because it is close to the genesis point of the storms shown in the upper-right panel of Figure 5. The two vertical lines correspond to important time points in the improvement of storm observations: first in 1945 when plane-based observations began, and second in 1966 when satellite-based tracking began [26]. It is evident that once the improved data from satellites became available, there is a close correspondence between SST and storm occurrence. We plan to exploit these, and more sophisticated, temporal and spatial relationships. In particular, our formal models for *conditional density estimation (CDE)* will incorporate SST (and similar variables) by transforming the available SST data from time series defined at each position on the ocean into time series defined at each position in diffusion space. There is a natural way of doing this, as each point in diffusion space corresponds to a track on the ocean: simply average SST over that track for each time point. The result will be a quantity which is localized both in space and time, and ready for direct comparison with the estimated TC density.

3.1. Orthogonal Series Density Estimation

The majority of work on methods for CDE has focused on the Nadaraya-Watson conditional kernel smoother first proposed by Rosenblatt [27], in which the conditional density is estimated as the ratio of the kernel density estimates for the joint density of the response and the predictors and the marginal density of the predictors [28, 29, 30, 31, 32]; however, that form of CDE is not as amenable to a high-dimensional response as orthogonal series conditional density estimation in which the basis is adapted to the data.

²From the Smith-Reynolds Extended Reconstructed Sea Surface Temperature Data Set [25].

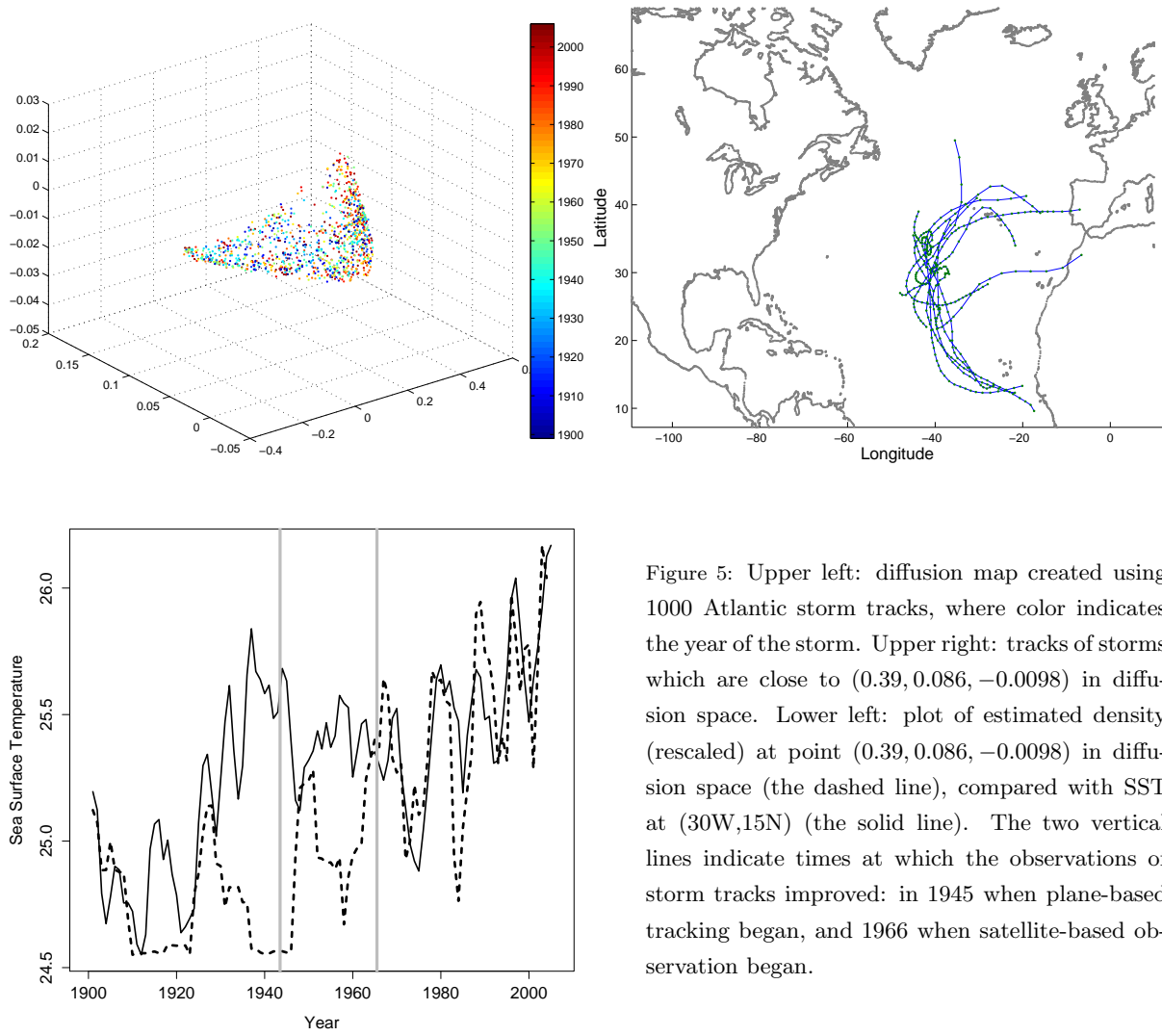


Figure 5: Upper left: diffusion map created using 1000 Atlantic storm tracks, where color indicates the year of the storm. Upper right: tracks of storms which are close to $(0.39, 0.086, -0.0098)$ in diffusion space. Lower left: plot of estimated density (rescaled) at point $(0.39, 0.086, -0.0098)$ in diffusion space (the dashed line), compared with SST at $(30W, 15N)$ (the solid line). The two vertical lines indicate times at which the observations of storm tracks improved: in 1945 when plane-based tracking began, and 1966 when satellite-based observation began.

Orthogonal series density estimation is motivated by the decomposition of a density as $f_Z(z) = \sum_{i=0}^{\infty} \theta_i \varphi_i(z)$ for an orthonormal series $\{\varphi_i : i \in \mathbb{Z}^+ \}$; the density is then estimated as $\hat{f}_Z(z) = \sum_{i=0}^k \hat{\theta}_i \varphi_i(z)$. Typically, noting that $E(\varphi_i(z)) = \int \left(\sum_{j=0}^{\infty} \theta_j \varphi_j(z) \right) \varphi_i(z) dz = \theta_i$, one estimates the Fourier coefficients as $\hat{\theta}_i = \frac{1}{n} \sum_{j=1}^n \varphi_i(z_j)$. The density estimation problem then is reduced to one of choosing the k that achieves the best bias-variance tradeoff.

In the conditional case, the series that we want to estimate is

$$f(y|x) = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \theta_{i,j} \varphi_{i,j}(x, y). \quad (11)$$

While this Fourier coefficient estimation scheme described above is trivially extended to the multivariate case, it cannot be neatly ported to the conditional case as the data points (x_i, y_i) are not samples from the conditional density $f_{Y|X}$ but from the joint density $f_{Y,X}$. But Efremovich [33] illustrates how the conditional case can be cast as an expectation with respect to the joint density:

$$\theta_{i,j} = \int \int f_{Y|X}(y|x) \varphi_{i,j}(x, y) dx dy = \int \int \frac{f_{Y,X}(y, x)}{f_X(x)} \varphi_{i,j}(x, y) dx dy = E \left(\frac{\varphi_{i,j}(X, Y)}{f_X(X)} \right). \quad (12)$$

Thus the conditional Fourier coefficients can be estimated as $\hat{\theta}_{i,j} = \frac{1}{n} \sum_{k=1}^n \frac{\varphi_{i,j}(x_k, y_k)}{f_X(x_k)}$.

This is relatively straightforward, yet we have not yet addressed the high dimension of the response y . The dimension reduction in this approach comes from using the orthonormal basis provided by the diffusion map, as it is *adapted to the intrinsic geometry of the data*. Specifically, if we think of $\varphi_{i,j}(x, y)$ as a tensor-product basis bifurcated by the predictor and response – $\varphi_{i,j}(x, y) = \phi_i(x)\lambda_j(y)$ – then the eigenfunctions estimated by the eigenvectors of the transition matrix P make a natural candidate for the response component of φ , i.e. $\lambda(y) = \psi(y)$.

In Richards et al. [7], this idea was successfully applied to a high-dimensional linear regression model using diffusion maps, in which the estimated eigenfunctions were used in the orthogonal series estimate of the regression function $r(\mathbf{x})$. (This worked well, because the $\psi(\mathbf{x})$ are sample approximations to smooth basis functions supported on the data (see [3]), and the relationship between the response and covariates (after reparameterization via the diffusion map) was sufficiently smooth. A more sophisticated basis may be required in this case due to the complex spatial variation in TCs; we are therefore also considering multi-scale bases based on a variation of the *treelet expansion* described in Lee et al. [34].) Girolami [35] employed a similar method for *unconditional* orthogonal series density estimation from kernel PCA; their basis is derived from the eigendecomposition of the Gram matrix.

The methods outlined in this paper provide a promising framework to address important scientific questions regarding the behavior of tropical cyclones. Using an approach to SCA, dimensionality reduction can be achieved without significant loss of important scientific information. The low-dimensional space yields a natural parameterization of the data, useful for constructing nonparametric density estimates and relating temporal and spatial variability in TCs to variations in other climate variables. In addition to the parameterization, the eigenvectors can be used as a basis for orthogonal series density estimation adapted to a high-dimensional setting.

References

- [1] D. W. Scott, Multivariate Density Estimation: Theory, Practice, and Visualization (Wiley Series in Probability and Statistics), Wiley-Interscience, ISBN 0471547700, 1992.
- [2] H. Liu, J. Lafferty, L. Wasserman, Sparse nonparametric density estimation in high dimensions using the rodeo, in: M. Meila, X. Shen (Eds.), Eleventh International Conference on Artificial Intelligence and Statistics (AISTATS), 2007.
- [3] A. B. Lee, L. Wasserman, Spectral Connectivity Analysis Submitted.
- [4] R. Coifman, S. Lafon, A. Lee, M. Maggioni, B. Nadler, F. Warner, S. Zucker, Geometric diffusions as a tool for harmonics analysis and structure definition of data: Diffusion maps, Proceedings of the National Academy of Sciences 102 (21) (2005) 7426–7431.
- [5] S. Lafon, A. B. Lee, Diffusion maps and coarse-graining: a unified framework for dimensionality reduction, graph partitioning, and data set parameterization, Pattern Analysis and Machine Intelligence, IEEE Transactions on 28 (9) (2006) 1393–1403.
- [6] M. Belkin, P. Niyogi, Laplacian Eigenmaps for dimensionality reduction and data representation, Neural Computation 15 (6) (2003) 1373–1396.
- [7] J. W. Richards, P. E. Freeman, A. B. Lee, C. M. Schafer, Exploring Low-Dimensional Structure in Astronomical Spectra, The Astrophysical Journal 691 (1) (2009) 32–42.
- [8] A. Y. Ng, M. I. Jordan, Y. Weiss, On Spectral Clustering: Analysis and an algorithm, in: Advances in Neural Information Processing Systems 14, MIT Press, 849–856, 2001.
- [9] U. von Luxburg, A Tutorial on Spectral Clustering, Statistics and Computing 17 (4) (2007) 395–416.
- [10] M. Belkin, P. Niyogi, Semi-supervised learning on Riemannian manifolds, in: Machine Learning, 209–239, 2004.
- [11] L. Xie, T. Yan, L. J. Pietrafesa, J. M. Morrison, T. Karl, Climatology and Interannual Variability of North Atlantic Hurricane Tracks, Journal of Climate 18 (24) (2005) 5370–5381.
- [12] HURDAT Database, URL http://www.aoml.noaa.gov/hrd/hurdat/Data_Storm.html, 2007.
- [13] T. Hall, S. Jewson, SST and North American Tropical Cyclone Landfall: A Statistical Modeling Study, [arXiv:0801.1013v1 \[physics.ao-ph\]](https://arxiv.org/abs/0801.1013v1), 2008.
- [14] T. M. Hall, S. Jewson, Statistical modelling of North Atlantic tropical cyclone tracks, Tellus (2007) 486–498.
- [15] J. Rumpf, H. Weindl, P. Hölpe, E. Rauch, V. Schmidt, Statistical modelling of tropical cyclone tracks, Mathematical Methods of Operations Research 66 (3) (2007) 475–490.
- [16] K. Emanuel, S. Ravela, E. Vivant, C. Risi, A statistical deterministic approach to hurricane risk assessment, Bulletin of the American Meteorological Society (2006) 299–312.
- [17] P. J. Vickery, P. F. Skerlj, L. A. Twisdale, Simulation of Hurricane Risk in the U.S. Using Empirical Track Model, Journal of Structural Engineering (2000) 1222–1237.
- [18] P. Arias, G. Randall, G. Sapiro, Connecting the Out-of-Sample and Pre-Image Problems in Kernel Methods, in: CVPR07, 1–8, 2007.
- [19] B. W. Silverman, Density Estimation for Statistics and Data Analysis, Chapman & Hall/CRC, ISBN 0412246201, 1986.
- [20] J. T. Y. Kwok, I. W. H. Tsang, The Pre-Image Problem in Kernel Methods, Neural Networks, IEEE Transactions on 15 (6) (2004) 1517–1525.
- [21] S. Mika, B. Schölkopf, A. J. Smola, K.-R. Müller, M. Scholz, G. Rätsch, Kernel PCA and De-Noising in Feature

- Spaces, in: M. S. Kearns, S. A. Solla, D. A. Cohn (Eds.), *Advances in Neural Information Processing Systems 11*, MIT Press, URL citeseer.ist.psu.edu/mika99kernel.html, 1999.
- [22] A. Justel, D. Peña, R. Zamar, A multivariate Kolmogorov-Smirnov test of goodness of fit, *Statistics and Probability Letters* 35 (1997) 251–259.
 - [23] P. Hall, N. Tajvidi, Permutation tests for equality of distributions in high-dimensional settings, *Biometrika* 89 (2) (2002) 359–374.
 - [24] P. Hall, N. E. Heckman, Estimating and depicting the structure of a distribution of random functions, *Biometrika* 89 (1) (2002) 145–158.
 - [25] T. M. Smith, R. W. Reynolds, T. Peterson, J. Lawrimore, Improvements to NOAA’s Historical Merged Land-Ocean Surface Temperature Analysis (1880–2006), *Journal of Climate* 21 (10) (2008) 2283–2296.
 - [26] G. Vecchi, T. R. Knutson, On Estimates of Historical North Atlantic Tropical Cyclone Activity, *Journal of Climate* 21 (14) (2008) 3580–3600.
 - [27] M. Rosenblatt, Conditional probability density and regression estimators, in: P. R. Krishnaiah (Ed.), *Multivariate Analysis II*, 1969.
 - [28] M. P. Holmes, A. G. Gray, C. L. Isbell, Jr., Fast Nonparametric Conditional Density Estimation, in: *The Learning Workshop (SNOWBIRD)*, 2007.
 - [29] J. G. D. Gooijer, D. Zerom, On Conditional Density Estimation, *Statistica Neerlandica* 57 (2) (2003) 159–176.
 - [30] R. J. Hyndman, D. M. Bashtannyk, G. K. Grunwald, Estimating and Visualizing Conditional Densities, *Journal of Computational and Graphical Statistics* 5 (4) (1996) 315–336.
 - [31] D. M. Bashtannyk, R. J. Hyndman, Bandwidth selection for kernel conditional density estimation, *Computational Statistics and Data Analysis* 36 (2001) 279–298.
 - [32] P. Hall, J. Racine, Q. Li, Cross-Validation and the Estimation of Conditional Probability Densities, *Journal of the American Statistical Association* 99 (468) (2004) 1015–1026.
 - [33] S. Efromovich, *Nonparametric Curve Estimation: Methods, Theory, and Applications*, Springer, 1999.
 - [34] A. B. Lee, B. Nadler, L. Wasserman, Treelets – An adaptive multi-scale basis for sparse unordered data, *Annals of Applied Statistics* 2 (2) (2008) 435–471.
 - [35] M. Girolami, Orthogonal series density estimation and the kernel eigenvalue problem, *Neural Computation* 14 (3) (2002) 669–688.